doi:10.3772/j.issn.2095-915x.2024.03.008

基于关键特征增强的金融长文本事件分类

开放科学（资源服务）标识码（OSID）

王洁[1] 李旭晖[1,2]

1. 武汉大学信息管理学院 武汉 430072；
2. 武汉大学大数据研究院 武汉 430072

摘要：[目的 / 意义] 为了解决长文本模型输入长度限制问题，通过抽取事件关键句和事件关键词，对长文本进行关键特征增强，以提高模型的特征表示能力。[方法 / 过程] 基于关键特征增强的模型，在原文的基础上利用 TextRank 算法抽取事件关键句，并利用 TF-IDF 算法抽取事件关键词，将二者作为关键特征对长文本进行特征增强，再利用 BERT 和 Self-Attention 模型进行特征的进一步提取，最后进行事件分类。[局限] 模型仅在金融领域事件分类上进行实验，可以考虑在其他领域内也进行实验并进一步验证模型效果。[结果 / 结论] 在金融长新闻事件分类数据集上，提出的模型准确率达到 88.40%，比基准模型提升了 2 个以上的百分点，表明了模型的有效性。

关键词：事件分类；长文本分类；关键特征；特征增强；自注意力机制

中图分类号：G35；TP391

Key-Features Enhanced Financial Long Text Event Classification

WANG Jie[1] LI Xuhui[1,2]

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Big Data Institute, Wuhan University, Wuhan 430072, China

Abstract: [Objective/Significance] In order to address the issue of length limitations in long-text models, this study enhances the feature representation capability of the model by extracting event-related key sentences and keywords from long text. [Methods/Processes] The key-features enhanced model utilizes the TextRank algorithm to extract key event sentences and the TF-IDF algorithm to extract event keywords from the original text. These key features are used to enhance the long text, and further feature extraction is performed using BERT and Self-Attention models, followed by event classification. [Limitations] The model in this study was only tested on event classification in the financial domain. It is recommended to conduct further experiments and verify the effectiveness of the model in other domains as well. [Results/Conclusions] On the financial long news event

基金项目 国家自然科学基金重点项目“基于知识关联的金融大数据价值分析、发现及协同创造机制”（91646206）；国家社科基金重大项目“文化遗产智慧数据资源建设与服务研究”（21&ZD334）。

作者简介 王洁（1998-），硕士研究生，主要研究方向为文本挖掘与金融文本分析；李旭晖（1975-），博士，副教授，主要研究方向为知识表示与自然语言处理，E-mail：lixuhui@whu.edu.cn。

引用格式 王洁 , 李旭晖 . 基于关键特征增强的金融长文本事件分类 [J]. 情报工程 , 2024, 10(3): 104-113.

classification dataset, the proposed model achieved an accuracy rate of 88.40%, outperforming other benchmark models by more than 2 percent, which demonstrates the superiority of the model.

Keywords: Event Classification; Long Text Classification; Key Features; Feature Enhancement; Self-Attention Mechanism

引言

金融新闻事件实时影响着市场参与者的情绪和决策，进而影响股票市场走势以及其他金融活动。对金融新闻事件进行分类有助于市场投资者快速掌握影响金融市场的重要事件，帮助其更好地进行分析和决策。在本研究中，一篇金融新闻即是对一个事件的报道，对金融事件进行分类即是根据新闻中报道的事件类别对文本进行分类。一方面，与传统的领域文本分类有所不同，事件分类主要关注的是领域内部的事件类别，根据文本中描述的具体事件将其分为某一事件类别，其分类更依赖事件要素相关词的区分。另一方面， 金融新闻文本中有许多篇幅较长的文章，要想对这类长文本进行分类，传统方法一般采用 TF-IDF 算法或 word-2vec 模型对文本进行向量化表示后再输入模型中训练，但其分类效果存在一定的局限性。以 BERT[1] 为代表的预训练语言模型在多项 NLP 任务上均超过了传统方法，但考虑到模型计算效率问题，BERT 模型的输入长度限制在 512 个 token[2]，这使得长文本无法完全输入 BERT 中进行表示。

基于以上问题和分析，针对长篇幅文本无法完全输入 BERT 模型的问题，本文提出从长文本中抽取出事件关键句的方法对原事件文本进行特征降维，达到将长文本中的关键信息抽取出来的目的。同时考虑到事件文本中的动词和名词往往更有可能携带关键的事件信息和事件要素，比如事件触发词和事件主客体等，本文采用 TF-IDF 算法从原长文本中抽取出关键词，再从中筛选出动词和名词，作为事件关键词。利用事件关键句和事件关键词中携带的关键信息作为原文的特征增强，有效地解决了长文本向量化表示的问题，并使模型学习到了更多关键信息，提升了金融新闻长文本事件分类的效果。本文的主要贡献如下：（1）针对文本过长无法完全输入模型的问题，采用关键句抽取的方法，将长文本进行特征降维，抽取出长文本中重要性更高的句子，作为原文的补充特征；（2）根据事件文本的特性，即动词和名词携带了更多事件信息，利用 TF-IDF 算法抽取出来的关键动词和名词作为事件关键词，对事件特征进行增强；（3）构建了金融领域内长文本事件分类的模型框架，将关键句特征和关键词特征作为原文的增强特征，提高了模型的特征表达能力。

1 相关工作

目前文本分类的常用方法主要可以分为基于传统机器学习的方法和基于深度学习的方法。早期文本分类任务使用的传统机器学习方法包括 NB[3]、SVM[4]、kNN[5] 等，而深度学习方法以 CNN 系列模型 [6] 和 RNN 及其变体 [7] 为主。自 Transformer 提出以来，预训练模型在 NLP 领域逐渐兴起。预训练模型分为以 BERT 为代表的自编码模型和以 GPT[8] 为代表的自回归模

doi:10.3772/j.issn.2095-915x.2024.03.008

型，GPT 主要用于文本生成任务，而 BERT 是一种基于双向 Transformer 构建的预训练模型，能够学习到文本的上下文特征表示。自提出以来，BERT 模型被广泛地应用到各 NLP 任务中，比如谌志群等 [9]、李颖 [10]、孙红等 [11] 均使用 BERT 模型进行文本向量化表示，再结合 CNN、LSTM、GRU 等模型进行下游任务，效果较传统方法也有显著提升。注意力机制因其能提升关键词的权重，也常被应用到文本分类相关研究中。Yang 等 [12] 在词和句级别分别应用注意力机制，使其能够在构建文档表示时关注更重要的内容。Shen 等 [13] 提出了一种用于语言理解的定向自注意力网络，使用注意力机制来学习句子嵌入。Basiri 等 [14] 提出了一种基于注意力机制的双向 CNN-RNN 深度模型用于情感分析。

针对长文本篇幅过长，无法完全输入 BERT 模型进行训练的问题，有相关研究通过句向量压缩或者分段等方式对长文本进行处理。叶瀚等 [15] 提出了一种句向量平均池化法以及注意力机制加权法对分类特征向量进行压缩编码进行长文本分类。卢玲等 [16] 将文本中的句子表示为段落向量，构建段落向量与文本类别的注意力模型计算句子的注意力，将句子注意力的均方差作为其对类别的贡献度，然后输入 CNN 中实现分类。鲍闯等 [17] 按照文本结构划分长文本，融合卷积最大池化特征向量和 BERT 句向量生成最终句向量，最后利用 Bi-LSTM 和注意力机制进行文本分类。长文本分类的相关研究较多采用对长文本切片进行文本表示，再对句向量进行平均等方式进行特征压缩，这种方式虽然能够利用到全局的文本信息，但是无法突出全文的关键信息，且对长文本的计算复杂度较高，对计算资源的要求也比较高。

事件分类和事件要素的抽取是构建事件知识图谱的基础，各大领域都有构建领域事件知识图谱的应用需求，比如自然灾难、体育赛事、历史事件等，金融事件也是其中一个重要的研究领域。Jacobs 等 [18] 构造了一个金融领域的英文新闻事件分类数据集，将金融事件划分为十大类别，包括买入评级、债务、股息、并购、盈利、季报、销售量、股份回购、目标价位、营业额，并在该数据集上分别采用线性 SVM 和 RNN-LSTM 进行实验。Jacobs 等 [19] 利用拥有 18 个类别和 64 个子类别的事件标注系统进行迭代标注，构建了一个英文金融事件分类数据集 SENTiVENT，并在 BERT 和 RoBERT 上进行了实验。Bhokare 等 [20] 对现有研究进行总结后将金融事件分为 11 类，并使用机器学习模型和 BERT 系列模型进行实验。已有研究大多基于英文新闻文本进行金融事件分类，而中文金融事件分类数据集几乎没有；同时，现有的金融事件分类研究重心在于金融事件类别的系统划分上，而没有对金融事件文本的特性进行分析，没有考虑如何提高金融事件分类的准确性。

2 关键特征增强的长文本事件分类

本文提出的基于关键特征增强的长文本事件分类模型整体结构如图 1 所示。整个模型由事件关键特征提取、事件关键特征增强、自注意力机制层、全连接层和 Softmax 分类层四个部分组成。整体流程为：（1）利用 TextRank 算法从原文中抽取事件关键句，再利用 TF-IDF

算法从原文中抽取出关键动词和名词作为事件关键词；（2）将原文、事件关键句、事件关键词分别输入 BERT 模型获得文本表示，再进行特征融合；（3）将增强后的特征输入自注意力层进行重要特征的进一步提取；（4）输入全连接层和 Softmax 激活函数进行事件分类。

2.1 事件关键特征提取

事件关键特征提取主要分为两部分：事件关键句提取和事件关键词提取。本文采用 TextRank 算法进行文本关键句抽取，得到的事件关键句能够表征长文本全局的主要信息，对文本长度过长无法输入模型的新闻文本进行特征补充。针对金融新闻事件分类任务，本文采用 TF-IDF 算法对全文词语计算重要性权重并排名，同时考虑到能够表征领域事件的词汇往往是领域名词和动词，本文仅筛选 TF-IDF 权重较高的动词和名词作为事件关键词。

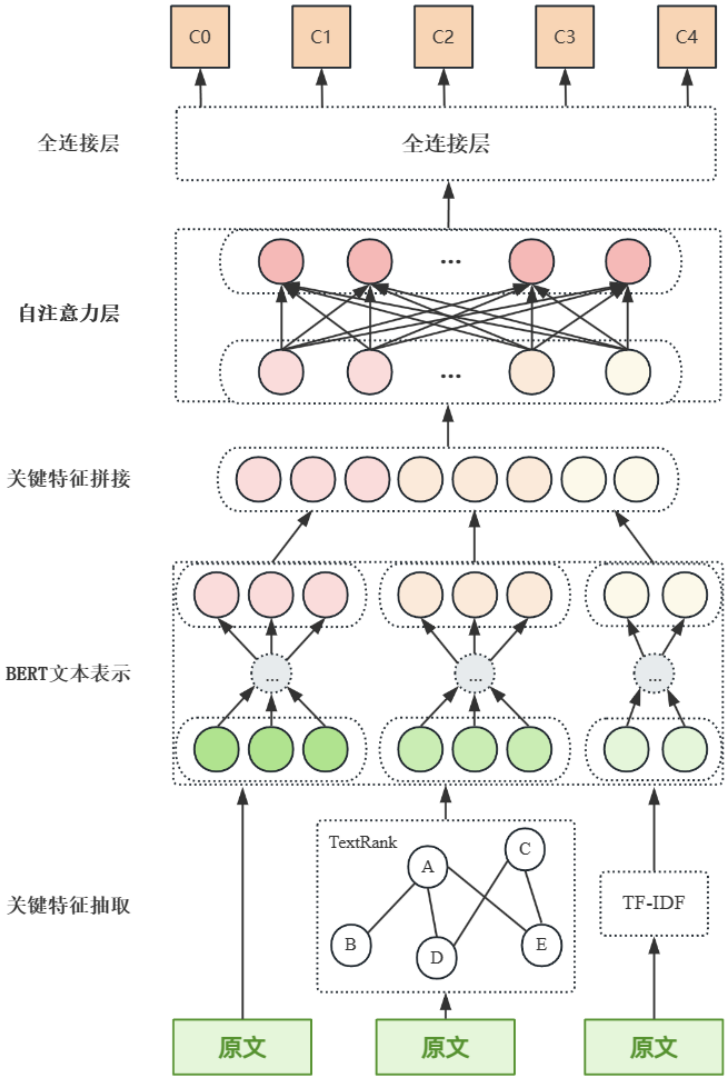

图 1　关键特征融合的金融长文本事件分类模型

doi:10.3772/j.issn.2095-915x.2024.03.008

2.1.1 事件关键句提取

TextRank[21] 算法源自 Google 提出的 PageRank 算法，本文使用的 TextRank 算法将长文本中的句子类比为 PageRank 算法中的网页，通过构建句子间的图结构关系并进行迭代计算，可以得到文本中比较重要的若干个句子作为新闻长文本的事件关键句。TextRank 的计算公式如下：

$$WS(v_i)=(1-d)+d\times\sum_{v_j\in In(v_i)}\frac{w_{ij}}{\sum_{v_k\in Out(v_j)}w_{jk}}WS(v_j) \tag{1}$$

其中，$WS(v_i)$ 表示节点 v_i 的权重，d 为平滑因子，表示在图结构中从一个节点转移到另一个节点的概率。$In(v_i)$ 表示节点 v_i 所有前驱节点的集合，$Out(v_j)$ 表示节点 v_j 所有后继节点的集合。w_{ij} 表示节点 v_i 和节点 v_j 之间的权重。

同时，本文采用以下公式来计算句子之间相似性作为节点之间边的权重：

$$Softmax(z_i)=\frac{\exp(z_i)}{\sum_j\exp(z_j)} \tag{2}$$

其中，$Sim(s_i, s_j)$ 表示句子 s_i 和句子 s_j 的相似度。w_k 表示两个句子的词集合中的第 k 个词，$|s_i|$ 和 $|s_j|$ 分别表示句子 s_i 和句子 s_j 的词的数量。句子 s_i 和 s_j 中相同词的个数越多，两个句子的相似度就越高；分母是对长度较长的句子进行一定程度的遏制，防止其因为长度优势而使句子间的相似度偏高。

2.1.2 事件关键词提取

事件关键词相对关键句而言更为短小精悍，核心语义信息更为集中，对事件文本进行关键词抽取能够从更细的粒度上对原文进行特征提取，提取出来的事件关键词能够进一步地对原文特征和事件关键句特征进行补充。同时，考虑到事件文本中的动词和名词往往携带了更多事件相关的信息，本文仅筛选 TF-IDF 算法排序靠前的动词和名词作为事件关键词。首先将原文经过分词、词性标注、去停用词等预处理步骤，然后采用 TF-IDF 算法对原文进行关键词重要性排序，并从中筛选出排名前 50 的动词和名词，得到事件关键词。

2.2 事件关键特征增强

自然语言文本需要转换成词向量表示，才能输入模型进行训练，本文采用 “bert-base-chinese” 模型进行文本的向量化表示。通过 TextRank 算法进行事件关键句抽取，得到长文本下的全局关键信息；同时，根据 TF-IDF 算法得到能够表征文档事件核心语义信息的事件关键词。至此，我们将原文 T、事件关键句 S 和事件关键词 W 分别输入 BERT 模型，得到词向量嵌入表示 F_t、F_s、F_w，然后将 F_s 和 F_w 拼接到原文的词向量 F_t 后面，得到进行特征增强后的文本表示 F_e。计算公式如下：

$$F_e=F_t\oplus F_s\oplus F_w \tag{3}$$

$\oplus$ 表示向量之间的拼接操作，上述公式将 F_s 和 F_w 拼接到 F_t 后面得到 F_e，得到经过事件关键特征增强之后的向量表示。

2.3 自注意力层特征提取

本步骤使用自注意力机制模型对得到关键特征增强之后的文本表示进行进一步的特征提取。具体而言，对于序列中的每个元素，自注意力机制计算该元素与其他元素之间的相似度，并将这些相似度归一化为注意力权重。然后，通过将每个元素与对应的注意力权重进行加权

求和，可以得到自注意力机制的输出。注意力机制的核心公式为：

$$Attention(Q,K,V)=Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

其中，Q 表示 query 矩阵，K 表示 key 矩阵，V 表示 value 矩阵，d_k 表示 k 矩阵的维度，除以 $\sqrt{d_k}$ 是为了使训练过程中的梯度值保持稳定。

2.4 全连接层和Softmax分类层

在经过上一个步骤的特征提取后，将得到的特征向量输入全连接层，再通过 Softmax 激活函数进行文本分类。全连接层在模型中的作用是可以有效地降低文本特征信息的损失。激活函数采用 Softmax 函数，将全连接层的特征向量输出作为 Softmax 函数的输入，然后进行金融新闻长文本的事件分类。Softmax 函数的具体计算公式如下：

$$Sim(s_i,s_j)=\frac{\left|\left\{w_k \mid w_k \in s_i \ \&\ w_k \in s_j\right\}\right|}{\log(|s_i|)+\log(|s_j|)} \tag{5}$$

3 实验与分析

3.1 实验数据准备

（1）实验环境

本文实验环境配置参数情况如表 1 所示。

表 1 实验环境配置参数

环境	配置参数
处理器	AMD Ryzen Threadripper 3970X 32-Core Processor 3.70 GHz
显卡	NVIDIA GeForce RTX 4090
内存	192G
编译器	PyCharm，Python 3.10

（2）数据准备

本文选取 5 类金融事件（人事变动、经营风险、证券市场风险、股权增减持、盈利能力），利用 python 网络爬虫方法，根据这 5 类事件标签从金融新闻资讯网站中获取实验数据，经过初步整理后得到每类 2000 条，共计 10000 条数据作为本次实验的数据集，并按照训练集：验证集：测试集为 8 ： 1 ： 1 的比例进行划分。同时，本文统计的文本数据长度分布情况如表 2 所示：

表 2 数据集信息

统计项	统计值
平均长度（字符）	1093
中位数（字符）	572
最短长度（字符）	68
最长长度（字符）	28276

由表 2 可见，金融新闻文本长度较长，最长达 28276 个字，平均长度为 1093 个字，中位数为 572 个字，这表明有一半以上的文本其长度超过了 BERT 模型所能处理的上限，超过的部分只能截断。因此，本文有必要对长新闻文本进行关键特征的抽取，以提高模型的事件分类效果。

3.2 实验设置与评价指标

（1）实验设置

实验参数设置如表 3 所示。

表 3 实验参数设置

参数	设定值
max_length	512
事件关键句数量	10
事件关键词数量	50
batch_size	32
epoch	15
optimizer	Adam

（2）评价指标

实验采用文本分类模型较为常用的查准率

doi:10.3772/j.issn.2095-915x.2024.03.008

（Precision）、查全率（Recall）、调和平均值（F1 值）和准确率（Accuracy）作为本文模型的评价指标。其计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{6}$$

$$R=\frac{TP}{TP+FN} \tag{7}$$

$$F1=\frac{2PR}{P+R} \tag{8}$$

$$A=\frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

3.3 实验结果与分析

3.3.1 文本输入长度对分类模型效果的影响

为了验证输入文本长度对分类模型效果的影响，本文以 BERT+Self-Attention 作为基础分类模型，选择文本输入长度为 128、256、512 分别进行对比实验，实验结果见图 2：

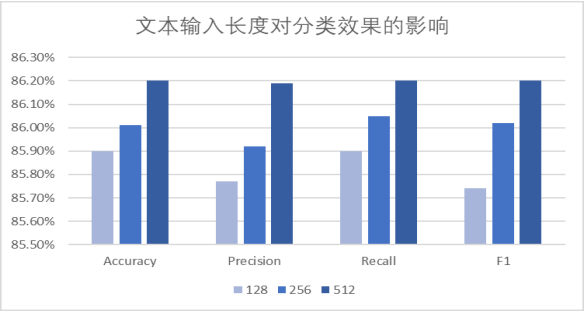

图 2 文本输入长度对比实验结果

由表中不同输入长度得到不同的模型效果，不难看出输入文本长度越长，模型能够从原文中获取到的信息越多，模型的分类效果也就越好。因此，针对 BERT 模型输入文本长度不超过 512 的限制，本文有必要通过关键特征抽取的方式，将全文的主要关键信息尽可能地输入模型中进行训练。

3.3.2 关键特征增强对分类模型效果的影响

在本部分对比实验中，为了探究关键句特征和关键词特征对分类模型的影响，本文以 BERT+Self-Attention 分类模型为基准，设置了 3 组对照实验，分别为：原文、原文＋关键句、原文＋关键词，实验组设置为：原文＋关键特征。同时为了验证关键句和关键词作为特征补充的思路的正确性，本文还设置了对比实验：关键句、关键词。实验结果对比如表 4 和图 3 所示：

表 4 关键特征增强的对比实验结果

模型	Accuracy	Precision	Recall	F1
原文	86.20%	86.19%	86.20%	86.08%
关键句	85.10%	85.17%	85.10%	85.02%
关键词	83.60%	83.52%	83.60%	83.39%
原文+关键句	87.20%	87.12%	87.20%	87.14%
原文+关键词	87.40%	87.46%	87.40%	87.30%
原文+关键特征	88.40%	88.33%	88.40%	88.28%

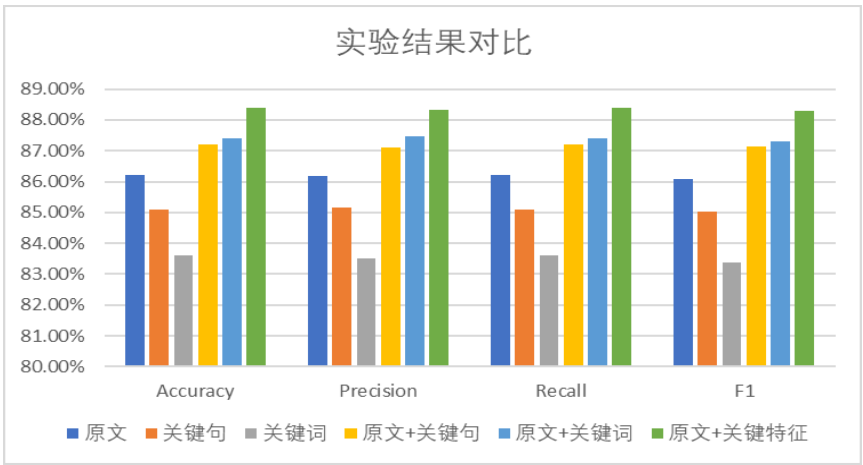

图 3 关键特征增强的对比实验结果

从表中的实验结果对比中可以看到，总体而言，本文提出的基于关键特征增强的长文本分类模型（原文＋关键特征）比基准模型（原文）的准确率提升了 2.2 个百分点，F1 值提升了 2.2 个百分点。从单个关键特征的角度来看，事件关键句特征增强方法（原文＋关键句）比（原文）的准确率和 F1 值均提升了 1 个百分点，表明本文提出的从长文本中抽取关键句进行特征增强的思路是可行且有效的，事件关键句中携带的文本信息能够对原文起到特征补充的作用；事件关键词特征增强方法（原文＋关键词）比（原文）的准确率和 F1 值均提升了 1.2 个百分点及以上，这表明本文抽取出来的事件关键词能够代表事件关键特征，能够携带事件相关信息。在与基准模型的对比之下，本文提出的将两个关键特征作为原文补充特征的模型效果最好，表明了该模型在领域内长文本事件分类任务上的可行性和有效性。

同时，观察到仅用关键句或关键词的分类效果并不如使用截断后的原文，推测可能的原因是事件关键句是从原文中按重要性排序抽取出来的句子，原本的语句顺序被打乱导致句子可能不具有语义连贯性；而仅用关键词输入模型的效果相对关键句的效果更差一些，原因是输入的文本为不连贯的词语，由词语本身携带的语义信息无法代表完整的文档信息，仅能突出部分关键特征。但将关键句和关键词作为原文的特征补充输入模型的效果均比原文有提升，表明以关键特征作为原文补充特征的思路是有效的，同时也表明 BERT 语言模型对文档全局语义信息的处理和利用是较为充分的，相比语义信息较为零散的单个句子和词语，其更擅长从全文中获取完整而连贯的语义信息。

表 5 展示了五个事件类别对应文档中提取出来的事件关键词，原文和事件关键句的详细信息由于篇幅过长不便展示。从表 5 中可以看到，人事变动事件类别文本中抽取出来的关键词包含“副行长”“行长”“监事长”等与企业人事职位变动相关的名词，也包括“担任”“辞职”“辞去”等职位变动相关的动词；经营风险类别中，包括“减值”“爆雷”“风险”等与公司经营风险相关的事件关键词；证券市场风险类别中包含“股价”“下跌”“收盘价”等与证券市场相关的专用词汇；股权增减持类别中包含了“减持”“股本”“持有”“股东”等与公司股东增减持事件相关的词；盈利能力类别中包括了“增速”“净利润”“扣非”“归母”“业绩”等描述企业业绩与盈利能力的关键词。这些事件关键词能够表达文本的核心语义信息，仅仅从几个简单的事件关键词便能勾勒出事件的大致轮廓，能够直观地反映金融事件类型。

3.3.3 与基准模型的对比

本文也与几个基准模型进行了对比。在传统机器学习模型中选择效果最好的线性 SVC 模型进行对比；在深度学习模型中则在 CNN 和 RNN 系列模型中选择文本分类最常用的 TextCNN 模型和 BiLSTM 模型进行对比；预训练语言模型中选择 BERT 分类模型进行对比。实验结果见表 6。

由表 6 可以看出，本文模型的事件分类效果最好。其中，线性 SVC 模型的事件分类准确率在 80.19% 左右，而深度学习分类模型的效果则明显高于传统的机器学习方法。TextCNN 模型和 Bi-LSTM 模型的准确率均在 86% 左右，

doi:10.3772/j.issn.2095-915x.2024.03.008

表 5　事件类别和事件关键词

事件类别	事件关键词
人事变动	银行 副行长 担任 董事会 行长 信息 辞职 充足率 截至 显示 首席 加入 高管 监事长 衷心感谢 股东 被执行人 辞呈 履历 职务 下降 原因 资本 记者 辞去 个人 强于 有限公司 资产 列入 发布公告 名单 董事 增长 做出 贡献 辞任 营收 人游 率为 拨备 换血 表示 资料 质量 负责 转任
经营风险	减值 净利润 爆雷 披露 公司 预告 业绩 投资者 基数 工业 集装箱 界面 增长 研究部 拐点 出现 景气 年报 风险 数据 历史 达到 年度 钢铁 应当 A股 等来 高箱 箱价 转跌 产组 新闻 满足 货箱 收益 商誉 结束 注意 来源 资产 下降 年度报告 股份 会计年度 要求 豁免 计提 超过 隐形 国际
证券市场风险	股价 下跌 产品 网络 企业级 网安 产业 领导者 企业 硬件 榜首 评为 新低 主营业务 触及 位居 收盘价 竞争力 客户 收盘 累计 简称 科技 截至 有限公司 发布 指数 提供 行业 历史 政府 继续 公司
股权增减持	减持 公司 普通股 股本 股份 显示 持有 股东 比例 内容 业务 交易 业务 范围 创意 视频 套现 均价 发布公告 产业链 互联网 儿童 定价 数字 净利润 资料 应用 从事 开发 公告 集中 报告 文化 方式 技术 产品 增长
盈利能力	增速 行业 装饰 净利润 创兴 扣非 归母 均值 排名 业绩 装修 东易 预减 名雕 预告 数据 精装 建筑 中值 资源 证券 下限 平均水平 上限 代码 时空 名称 经常性 损益 预增 低于 类型 公告 减少 来源 股份 年度 披露 公司 科技 名列 缩减 主营业务 指标 预喜 增亏 截至 收益 甲方 免责

表 6　基准模型对比实验结果

	accuracy	precision	recall	F1
线性SVC	80.19%	80.15%	79.58%	79.86%
TextCNN	86.05%	85.90%	86.05%	85.97%
BiLSTM	85.95%	85.93%	85.82%	85.90%
BERT	86.42%	86.38%	86.12%	86.32%
本文模型	88.40%	88.33%	88.40%	88.28%

比线性 SVC 模型高出 5 个百分点以上。BERT 模型准确率在 86.42% 左右，而本文基于关键特征增强的模型效果在 88.40% 左右，比 TextCNN、Bi-LSTM 模型和 BERT 模型均高出 2 个百分点左右，表明本文提出的模型在长文本事件分类任务上有更好的效果。

4　总结展望

现有文本分类研究关注领域之间的文本分类较多，而关注领域内部事件分类的相对较少，本文以金融领域为例，研究该领域下的长新闻事件分类问题。为了解决长文本无法完全输入模型的问题，并根据事件的特性捕获更多事件相关特征，本文提出基于关键特征增强的长文本事件分类模型，通过 TextRank 算法抽取长文本的事件关键句，再用 TF-IDF 算法抽取出表征事件信息的动词和名词作为事件关键词，经过 BERT 文本表示并对原文进行特征补充后，输入 SelfAttention 模型实现金融新闻长文本的事件分类任务。

虽然本文的关键特征增强模型有一定的创新性，也在金融领域长文本事件分类任务上比基准模型表现更好，但是本研究依然存在一定的不足和改进空间。本文的关键句抽取方法采用 TextRank 算法，其中句子节点之间的转移概率为句子间的相似度，而句子间的相似度计算方法可以进行改进以抽取出更准确的事件关键句；事件关键词抽取算法采用传统的 TF-IDF 算法，在抽取算法上也存在较大的改进空间。未来的工作将考虑在这两个方面进行更深入的

研究。

参 考 文 献

[1] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[2] DING M, ZHOU C, YANG H, et al. Cogltx: Applying bert to long texts[J]. Advances in Neural Information Processing Systems, 2020, 33: 12792-12804.

[3] 丁月 , 汪学明 . 基于改进特征加权的朴素贝叶斯分类算法 [J]. 计算机应用研究 , 2019, 36(12): 3597-3600, 3627.

[4] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[5] 丁正生 , 马春洁 . 改进词向量和 kNN 的中文文本分类算法 [J]. 现代电子技术 , 2022, 45(1): 100-103.

[6] KALCHBRENNER N, GREFENSTETTE E, BLUNSOM P. A Convolutional Neural Network for Modelling Sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 655-665.

[7] 贺波 , 马静 , 李驰 . 基于融合特征的商品文本分类方法研究 [J]. 情报理论与实践 , 2020, 43(11): 162-168.

[8] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. 2018.

[9] 谌志群 , 鞠婷 . 基于 BERT 和双向 LSTM 的微博评论倾向性分析研究 [J]. 情报理论与实践 , 2020, 43(8): 173.

[10] 李颖 . 基于 BERT-DPCNN 的垃圾弹幕识别改进及应用 [D]. 上海 : 上海师范大学 , 2020.

[11] 孙红 , 陈强越 . 融合 BERT 词嵌入和注意力机制的中文文本分类 [J]. 小型微型计算机系统 , 2022, 43(1): 22-26.

[12] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

[13] SHEN T, ZHOU T, LONG G, et al. Disan: Directional self-attention network for rnn/cnn-free language understanding[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1): 5446-5455.

[14] BASIRI M E, NEMATI S, ABDAR M, et al. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis[J]. Future Generation Computer Systems, 2021, 115: 279-294.

[15] 叶瀚 , 孙海春 , 李欣 , 等 . 融合注意力机制与句向量压缩的长文本分类模型 [J]. 数据分析与知识发现 , 2022, 6(6): 84-94.

[16] 卢玲 , 杨武 , 王远伦 , 等 . 结合注意力机制的长文本分类方法 [J]. 计算机应用 , 2018, 38(5): 1272-1277.

[17] 鲍闯 , 乔杰 , 李海斌 . 基于融合特征的长文本分类方法 [J]. 重庆理工大学学报 (自然科学), 2022, 36(9): 128-136.

[18] JACOBS G, LEFEVER E, HOSTE V. Economic event detection in company-specific news text[C]//1st Workshop on Economics and Natural Language Processing (ECONLP) at Meeting of the Association-for-Computational-Linguistics (ACL). Association for Computational Linguistics (ACL), 2018: 1-10.

[19] JACOBS G, HOSTE V. SENTiVENT: enabling supervised information extraction of company-specific events in economic and financial news[J]. Language Resources and Evaluation, 2022, 56(1): 225-257.

[20] BHOKARE P, SONAWANE A, SONAWANE S, et al. Detection and Classification of Financial Events from News Articles[J]. Iconic Research and Engineering Journals, 2023, 6(8): 171-180.

[21] MIHALCEA R, TARAU P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.